A41A-3002: Parameter Estimation in CLM4.5 Using Surrogate Model Based Global Optimization

Juliane Müller[(1)], Rajendra Paudel[(2)], Christine Shoemaker[(2)], Joshua Woodbury[(2)], Yilun Wang[(2)], Natalie Mahowald[(2)]
(1) Lawrence Berkeley National Lab (2) Cornell University

Problem Description

The Community Land Model (CLM) of the Community Earth System Models contains a module to estimate CH_4 emissions from natural wetlands and rice paddies. However, there are large discrepancies between the CLM predictions and the observations. Our goal is to use surrogate optimization in order to tune the model parameters such that the model predictions and the observations are in better agreement.

The difficulty of solving the optimization problem lies in the objective function which has the following properties:

1. computationally expensive
2. black-box
3. multi-modal/possibly not identifiable

Goal: Find a near optimal solution (set of tuning parameters) within very few expensive objective function evaluations such that observations and model predictions match better.

Surrogate Optimization has been shown to be very efficient for computationally expensive model calibration since it obtains accurate answers with generally fewer model evaluations than competing methods. To our knowledge this is the first application of a surrogate optimization method to calibrate a global climate model.

Mathematical Problem Formulation

Our goal is to minimize the sum of weighted root mean squared errors (RMSE):

$$\min \quad f(x) = \sum_{i=1}^{M} w_i r_i \qquad (1)$$

$$\text{such that} \quad -\infty < x_k^l \leq x_k \leq x_k^u < \infty, k = 1, \ldots, d \qquad (2)$$

where

1. M is the number of locations where we have observations
2. $r_i = \sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}[O_{i,j} - S_{i,j}(x)]^2}, i = 1, \ldots, M$ is the RMSE at location i
3. N_i is the number of observations at location i
4. $O_{i,j}$ is the jth observation at the ith location
5. $S_{i,j}(x)$ is the jth model prediction at the ith location given the parameter vector x
6. w_i is a weight that is computed based on the total CH_4 emissions at location i

Simulation Model Description

- We used CLM4.5bgc (latest CLM version with improved biogeochemistry).
- We used a mechanistic CH_4 emission model [1,6].
- CLM4.5bgc simulates the physical and biogeochemical processes regulating terrestrial CH_4 fluxes (CH_4 production, oxidation, CH_4 and O_2 transport through aerenchyma of wetland plants, ebullition, and CH_4 and O_2 diffusion through soil).
- CLM4.5bgc includes constraints on CH_4 emissions such as the effects of redox potential and soil pH.
- CLM4.5bgc can simulate satellite derived inundation fraction.

Model Parameters and Observation Sites

- CLM4.5bgc has 21 CH_4-related parameters .
- We used parameter bounds (x_k^l and x_k^u) based on literature values.
- We used sensitivity analyses to determine the most important parameters.
- We found 11 parameters that were important for almost all observation sites.
- We used observations from 6 natural wetland and 10 rice paddy sites around the globe.
- The number of observations at each site ranged from 10 to 79 collected over 1 to 3 years.

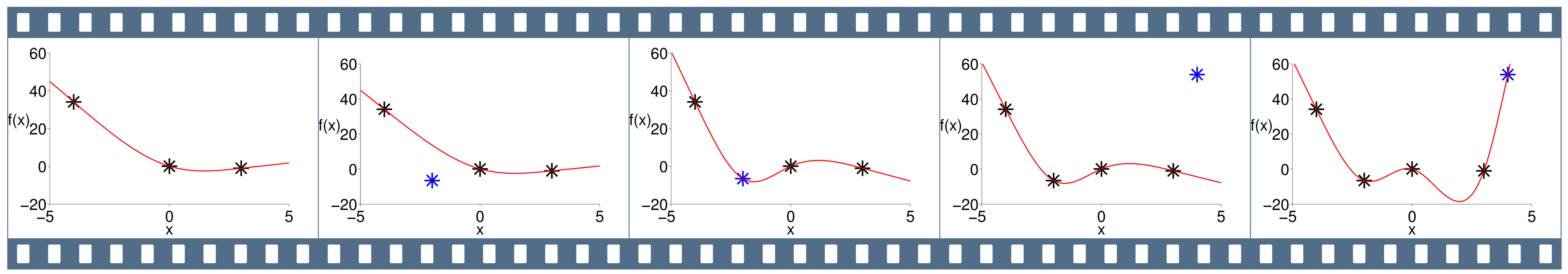
Surrogate Optimization

In surrogate optimization, we use a computationally cheap approximation $s(x)$ of the expensive objective function $f(x)$ in order to predict function values at unsampled points:

$$f(x) = s(x) + e(x).$$

A general surrogate algorithm works as follows (see also [2, 3, 5]):

Step 1: Create an initial experimental design and evaluate $f(x)$ at the selected points. Fit the surrogate $s(x)$ to the data.
Step 2: Use $s(x)$ to select a new evaluation point x_{new} and compute $f(x_{\text{new}})$.
Step 3: If the stopping criterion is not satisfied, update $s(x)$ with the new data and go to Step 2. Otherwise, stop.



Step 1 Step 2 Step 3 Step 2 Step 3

We use a cubic radial basis function interpolant as surrogate

$$s(x) = \sum_{l=1}^{n} \lambda_l \phi(\|x - x_l\|) + p(x)$$

where n is the number of already evaluated points, x_l is the lth already evaluated point, $\phi(r) = r^3$ is the cubic radial basis function, $\|\cdot\|$ is the Euclidean norm, and $p(x) = b_0 + b^T x$ is a polynomial tail. The parameters λ_l, b_0, and b are determined by solving a linear system of equations:

$$\begin{bmatrix}\Phi & P\\ P^T & 0\end{bmatrix}\begin{bmatrix}\lambda\\ c\end{bmatrix} = \begin{bmatrix}F\\ 0\end{bmatrix}, \text{ where } P = \begin{bmatrix}x_1^T & 1\\ x_2^T & 1\\ \vdots & \vdots\\ x_n^T & 1\end{bmatrix}, \lambda = \begin{bmatrix}\lambda_1\\ \lambda_2\\ \vdots\\ \lambda_n\end{bmatrix}, c = \begin{bmatrix}b_1\\ b_2\\ \vdots\\ b_d\\ b_0\end{bmatrix}, F = \begin{bmatrix}f(x_1)\\ f(x_2)\\ \vdots\\ f(x_n)\end{bmatrix}, \text{ and } \Phi_{\iota\nu} = \phi(\|x_\iota - x_\nu\|), \iota, \nu = 1, \ldots, n. \qquad (3)$$

The matrix in equation (3) is invertible if and only if $\text{rank}(P) = d + 1$ [4].

Numerical Experiments with Pseudo Data

Pseudo data case to verify that the algorithm works:

- We created pseudo data by running CLM4.5bgc with default parameters for all observation sites.
- We recorded the model's CH_4 output for the same dates at which we have observations (=pseudo observations).
- We perturbed the parameters and tried to find the default parameters by surrogate optimization.
- The globally optimal objective function value in the pseudo data case is 0.

Results:

- We did three trials. The final objective function values were close to 0.
- Initial objective function values were larger than 50, within less than 100 evaluations we could decrease that value to less than 10 in all trials.
- We found that the problem has several local optima with very similar objective function values. Hence, the problem is probably not identifiable.

Acknowledgements and Contact

The authors want to acknowledge the funding sources DOE SciDAC DE-SC0006791, NSF 1049031, NSF 1049033, and NSF CISE 1116298. The first author also wants to acknowledge partial support by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under contract number DE-AC02005CH11231.

- Juliane Müller: juliane.mueller2901@gmail.com (corresponding author)
- Rajendra Paudel: rp336@cornell.edu
- Chris Shoemaker: cas12@cornell.edu
- Natalie Mahowald: mahowald@cornell.edu

Numerical Experiments with Real Data (1)

Real data case to improve the model's parameters:

- We include the default parameter values in the initial experimental design.
- We use the real observations to compute the objective function values corresponding to the parameter vectors for which we run CLM4.5bgc.
- The globally optimal objective function value in the real data case should be lower than the objective function value for the default parameters.

Results:

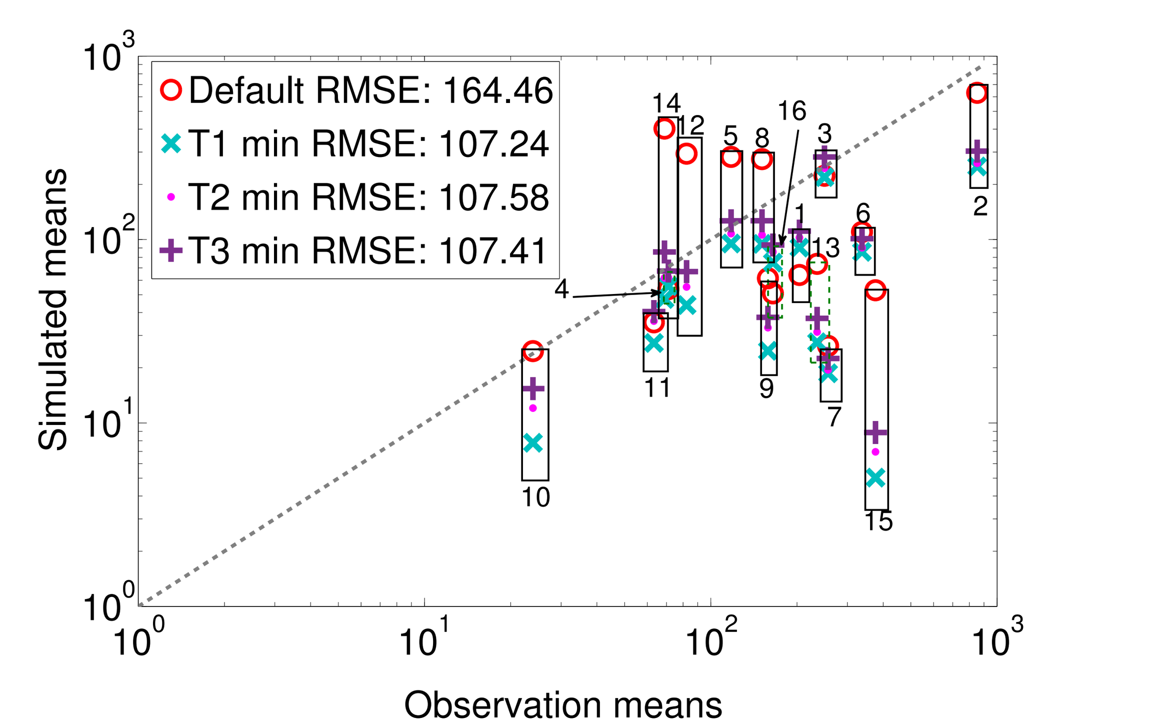


Fig. 1: Scatter plot of CH_4 observation means versus simulated means with default and optimized parameters for all observation sites. Northern: 1, 16; (sub)tropical: 5, 8, 12, 14, 15; temperate: 2, 3, 4, 6, 7, 9, 10, 11, 13. The closer the markers are to the dashed line, the better observations and simulations agree.

Numerical Experiments with Real Data (2)

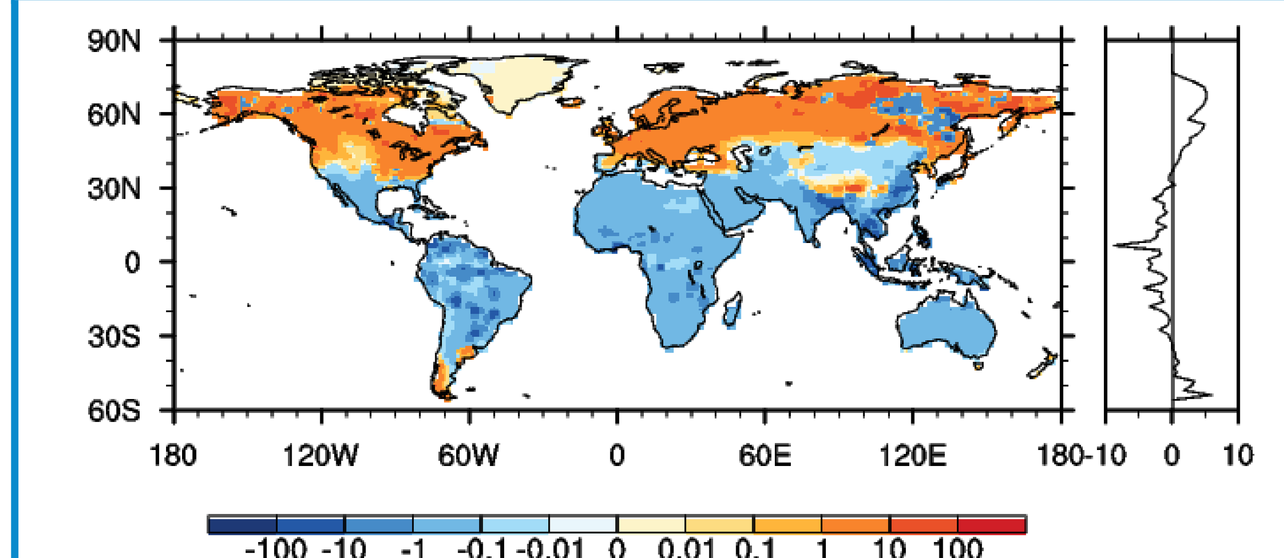


Fig. 2: Difference between predicted CH_4 emissions (mg $CH_4/m^2/d$) with optimized parameters and default parameters. With the optimized parameters, the CH_4 emission predictions in the northern regions are larger than for the default parameters. For the tropics, the predictions with the optimized parameters are lower than when using the default values.

Numerical Experiments with Real Data (3)

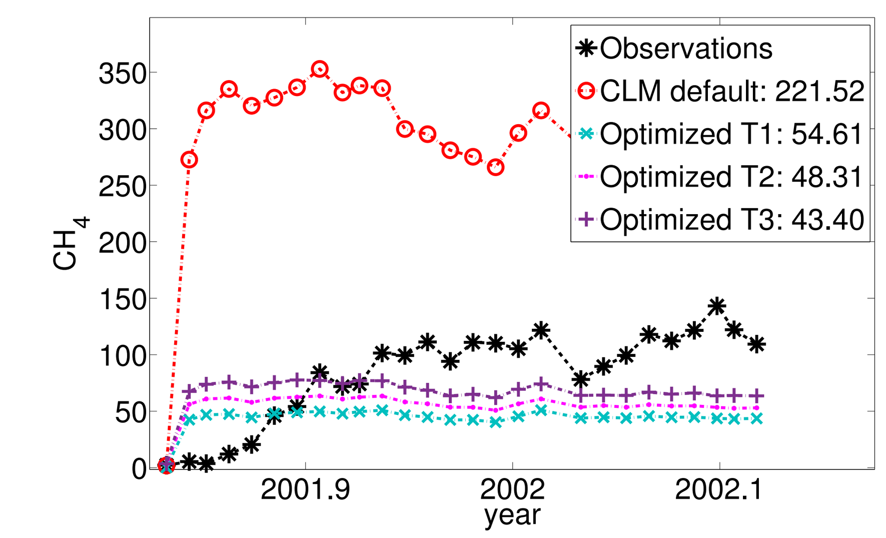


Fig. 3: Observation data, predictions with default and optimized parameters, and corresponding RMSE's for Central Java, Indonesia, observation site.

Conclusions

- We tuned the CH_4 related parameters in CLM4.5bgc using surrogate optimization to achieve a better fit of observations and model predictions.
- We assessed the effectiveness of the surrogate algorithm by using pseudo data.
- For the real data, the algorithm reduced the objective function value significantly as compared to the default value.
- We found that the objective function landscape is multimodal, hence the problem is probably not identifiable.
- The total global CH_4 emissions using the optimized parameters does not change significantly, but the distribution of CH_4 emissions between latitudes changed.
- The observation data drives the model to predict more emissions in the northern latitudes and less in the tropics.

References

[1] Meng et al. Sensitivity of wetland methane emissions to model assumptions: application and model testing against site observations, Biogeosciences, 9, 2012.
[2] Müller and Piché. Mixture surrogate models based on Dempster-Shafer theory for global optimization problems, Journal of Global Optimization, 51, 2011.
[3] Müller and Shoemaker. Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems, Journal of Global Optimization, 60, 2014.
[4] Powell. The theory of radial basis function approximation in 1990, Advances in Numerical Analysis, vol. 2. Oxford University Press, Oxford, 1992.
[5] Regis and Shoemaker. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization, Engineering Optimization, 45, 2013.
[6] Riley et al. Barriers to predicting changes in global terrestrial methane fluxes: analyses using CLM4Me, a methane biogeochemistry model integrated in CESM, Biogeosciences Discussions, 2011.